



# Pengenalan Pola

Klasterisasi Data

PTIHK - 2014



# Course Contents

1

Konsep Dasar

2

Tahapan Proses Klasterisasi

3

Ukuran Kemiripan Data

4

Algoritma Klasterisasi



## Konsep Dasar

- Klusterisasi Data, atau *Data Clustering* (atau *Clustering*), juga disebut sebagai analisis klaster, analisis segmentasi, analisis taxonomi, atau *unsupervised classification*
- Metode yang digunakan untuk membangun grup dari objek-objek, atau klaster-klaster, dimana objek-objek dalam satu kluster tertentu memiliki kesamaan ciri yang tinggi dan objek-objek pada kluster yang berbeda memiliki kesamaan ciri yang rendah



## Konsep Dasar

- Tujuan dari klasterisasi data adalah mengelompokkan data yang memiliki kesamaan ciri dan memisahkan data ke dalam klaster yang berbeda untuk objek-objek yang memiliki ciri yang berbeda
- Berbeda dengan klasifikasi, yang memiliki klas yang telah didefinisikan sebelumnya. Dalam klasterisasi, klaster akan terbentuk sendiri berdasarkan ciri objek yang dimiliki dan kriteria pengelompokan yang telah ditentukan.



## Konsep Dasar

- Untuk menunjukkan klasterisasi dari sekumpulan data, suatu kriteria pengelompokan haruslah ditentukan sebelumnya.
- Perbedaan kriteria pengelompokan akan memberikan dampak perbedaan klaster juga



## » Contoh

- Dua klaster dengan kriteria “How mammals bear their progeny”

Blue shark,  
sheep, cat,  
dog

Lizard, sparrow,  
viper, seagull, gold  
fish, frog, red  
mullet

- Dua klaster dengan kriteria “Existence of lungs”

Gold fish, red  
mullet, blue  
shark

Sheep, sparrow,  
dog, cat, seagull,  
lizard, frog, viper




# Tahapan Klasterisasi

1. Feature Selection
  - Penentuan informasi fitur yang digunakan
2. Proximity Measure
  - Tahap kuantifikasi item kemiripan data
3. Clustering Criterion
  - Penentuan fungsi pembobotan / tipe aturan
4. Clustering Algorithm
  - Metode klaster berdasarkan ukuran kemiripan data dan kriteria klasterisasi
5. Validation of the Result
6. Interpretation of the Result



## » Proximity Measure

- Kemiripan data memiliki peranan yang sangat penting dalam proses analisis klaster
  - Pada berbagai literatur tentang *clustering*, ukuran kemiripan (*similarity measures*), koefisien kemiripan (*similarity coefficients*), ukuran ketidakmiripan (*dissimilarity measures*), atau jarak (*distances*) digunakan untuk mendeskripsikan nilai kuantitatif dari kemiripan atau ketidakmiripan dari dua titik atau dua klaster
- 



## Proximity Measure

- Koefisien kemiripan data mengindikasikan **kekuatan hubungan** antar dua data
- Semakin banyak kemiripan satu sama lain, semakin besar koefisien kesamaan
- Misal  $x = (x_1, x_2, \dots, x_d)$  dan  $y = (y_1, y_2, \dots, y_d)$  dua data titik pada d-dimensi. Koefisien kemiripan data antara x dan y merupakan fungsi jarak dari nilai atribut-atribut nya

$$s(\mathbf{x}, \mathbf{y}) = s(x_1, x_2, \dots, x_d, y_1, y_2, \dots, y_d).$$

# Proximity Measure

- Pemilihan jarak pada aplikasi clustering adalah sangat penting, dan pilihan yang terbaik sering diperoleh melalui pengalaman, kemampuan, pengetahuan, dan keberuntungan.
- Pengukuran Data
  - Numerik
    - Euclidean Distance
    - Manhattan Distance
    - Maximum Distance
    - Minkowski Distance
    - Mahalanobis Distance
    - Average Distance
  - Kategorikal
    - Simple Matching Distance

## » Euclidean Distance

- Euclidean distance merupakan pengukuran jarak yang paling umum digunakan pada data numerik.
- Untuk dua data titik  $\mathbf{x}$  dan  $\mathbf{y}$  dalam  $d$ -ruang dimensi, Euclidean distance antara titik tersebut didefinisikan sebagai berikut :

$$d_{euc}(\mathbf{x}, \mathbf{y}) = \left[ \sum_{j=1}^d (x_j - y_j)^2 \right]^{\frac{1}{2}} = [(\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^T]^{\frac{1}{2}}$$

dimana  $x_j$  dan  $y_j$  masing-masing adalah nilai dari atribut ke- $j$  dari  $\mathbf{x}$  dan  $\mathbf{y}$

## » Manhattan Distance

- Manhattan distance disebut juga sebagai “*city block distance*” merupakan jumlah jarak dari semua attribute.
- Untuk dua data titik **x** dan **y** dalam **d**-ruang dimensi, Manhattan distance antara titik tersebut didefinisikan sebagai berikut :

$$d_{man}(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^d |x_k - y_k|.$$



## Maximum Distance

- Maximum distance disebut juga sebagai “*sup distance*”. Didefinisikan sebagai nilai maximum dari jarak atribut-atribut nya
- Untuk dua data titik **x** dan **y** dalam **d**-ruang dimensi, Maximum distance antara titik tersebut didefinisikan sebagai berikut :

$$d_{max}(x, y) = \max_{1 \leq k \leq d} |x_k - y_k|.$$



# Minkowski Distance

- Euclidean distance, Manhattan distance, dan maximum distance merupakan tiga kasus khusus dari Minkowski distance yang didefinisikan sebagai berikut :

$$d_{min}(\mathbf{x}, \mathbf{y}) = \left( \sum_{j=1}^d |x_j - y_j|^r \right)^{\frac{1}{r}}, \quad r \geq 1.$$

$r$  disebut sebagai order dari Minkowski distance. Jika  $r = 2, 1$ , and  $\infty$ , maka formulasi jarak tersebut masing-masing adalah Euclidean distance, Manhattan distance, and maximum distance

## ➤ Mahalanobis Distance

- Mahalanobis distance dapat mengurangi distorsi (penyimpangan) jarak yang disebabkan oleh kombinasi linier dari atribut.
- Mahalanobis distance didefinisikan sebagai berikut:

$$d_{mah}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y}) \Sigma^{-1} (\mathbf{x} - \mathbf{y})^T},$$

dimana  $\Sigma$  adalah matriks covariance dari data. Oleh sebab itu, jarak ini mengaplikasikan skema bobot terhadap data

## ➤ Average Distance

- Dua titik data dalam Euclidean distance tanpa nilai attribute biasanya memiliki jarak lebih kecil daripada pasangan data yang mengandung nilai.
- Pada kasus tersebut, average distance dikembangkan untuk mengatasinya

$$d_{ave}(\mathbf{x}, \mathbf{y}) = \left( \frac{1}{d} \sum_{j=1}^d (x_j - y_j)^2 \right)^{\frac{1}{2}}.$$

- Average distance merupakan hasil modifikasi dari Euclidean distance

## Simple Matching Distance

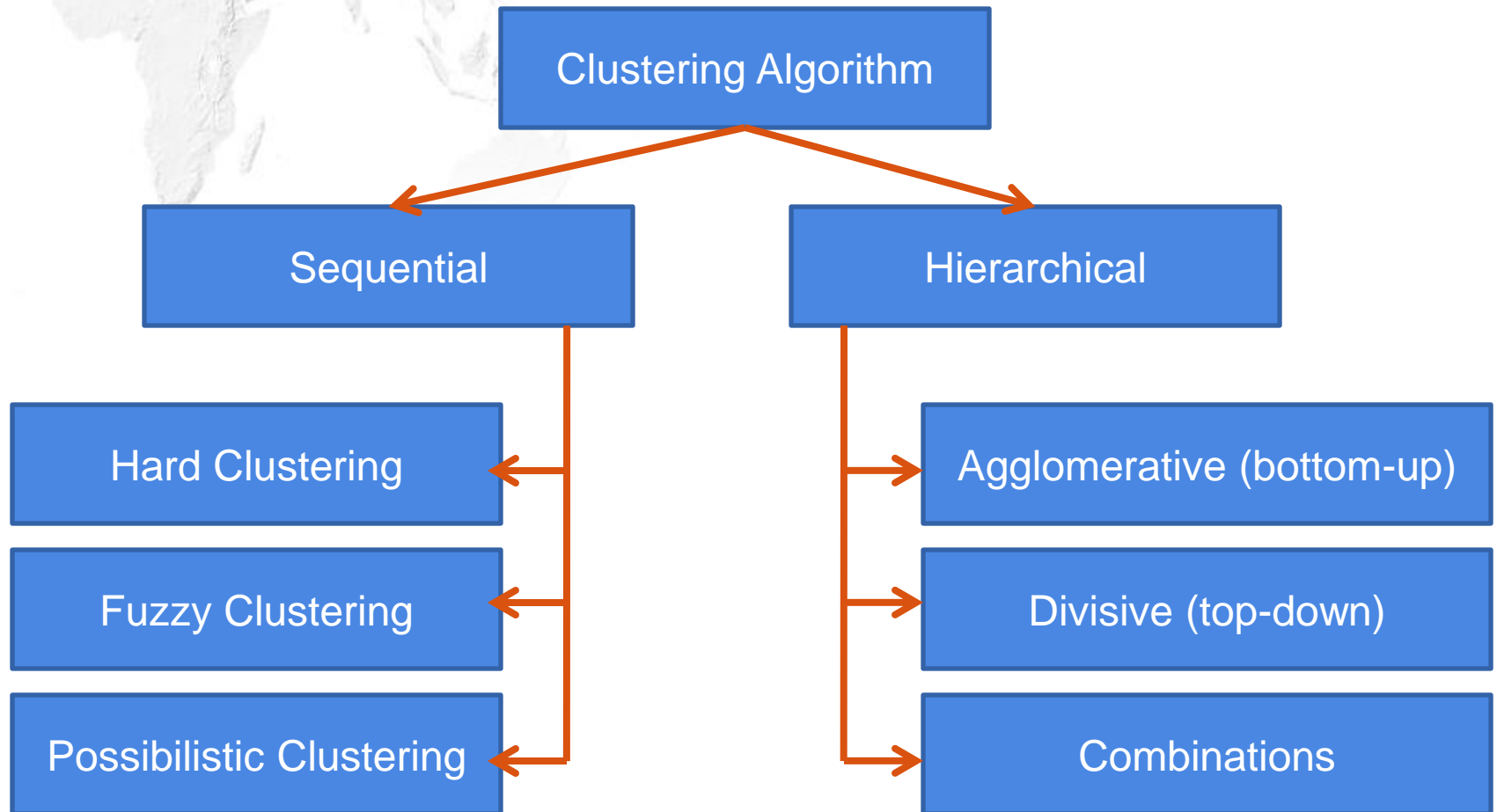
- Misal  $x$  and  $y$  dua nilai data kategorikal. Maka simple matching distance antara  $x$  dan  $y$  didefinisikan oleh:
- Misal  $x$  dan  $y$  dua objek data kategorikal dideskripsikan oleh  $d$  atribut kategorikal. Maka pengukuran kemiripan antara  $x$  dan  $y$  menggunakan simple matching distance didefinisikan oleh:

$$\delta(x, y) = \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{if } x \neq y. \end{cases}$$

$$d_{sim}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^d \delta(x_j, y_j).$$



# Clustering Algorithm





# » Clustering Algorithm

- **Sequential:** merupakan single clustering. One or few sequential passes on the data.
- **Hierarchical:** merupakan sekuen dari nested clusterings.
- **Hard clustering:** setiap titik data merupakan anggota dari satu klaster secara eksklusif
- **Fuzzy clustering:** setiap titik data merupakan anggota lebih dari satu klaster secara simultan
- **Possibilistic clustering:** klasterisasi yang didasarkan pada *possibility* suatu titik data terhadap klaster



# Hard Clustering Algorithm

- Hard Clustering
  - Basic hard clustering algorithms (e.g.,  $k$ -means)
  - $k$ -medoids algorithms
  - Mixture decomposition
  - Branch and bound
  - Simulated annealing
  - Deterministic annealing
  - Boundary detection
  - Mode seeking
  - Genetic clustering algorithms



# Fuzzy Clustering Algorithm

- Fuzzy Clustering
  - Fuzzy  $k$ -means
  - Fuzzy  $k$ -modes
  - Fuzzy c-means



# Hierarchical Clustering Algorithm

- Agglomerative Hierarchical
  - Graph method
    - Single-link method
    - Complete-link method
    - Group average method
    - Weighted group average method
  - Geometric method
    - Ward's method
    - Centroid method
    - Median method
- Divisive Hierarchical → kebalikan dari agglomerative



# Thank You !

afif.supianto@ub.ac.id  
081 331 834 734 / 088 160 127 40